Computer Science and Application 计算机科学与应用, 2021, 11(1), 121-132
Published Online January 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.111014

Hans汉斯

车企舆情正负面识别与预测

武　壮

云南财经大学统计与数学学院，云南　昆明
Email: 1148322387@qq.com

收稿日期：2020年12月25日；录用日期：2021年1月19日；发布日期：2021年1月26日

摘　要

随着科技的不断进步，人们生活越来越好，车辆普及度逐渐提高，人们也越来越关注车辆带给他们的体验。而对于汽车企业而言，汽车安全直接关乎客户的生命安全，人们对于车企舆情的正负面也有着更高的关注度和敏感性，舆情处理难度只会更大。如果负面舆情不能及时处理，车企将面临着重大的舆论压力，而且事后进行处理时也会耗费大量的资源和财力。由于产品大多生产规模庞大，多方利益纠缠，车企的舆情系统往往比其他企业有更高的舆情要求，所以对于汽车企业而言，舆情的识别与预测起着很重要的作用。本文通过建立朴素贝叶斯模型对车企舆情正负面进行识别与预测，在有效处理数据的基础上，利用给出的训练集数据建立模型，用测试集数据对模型的合理性和科学性进行评估验证。研究表明，本文所采取的车企舆情识别与预测模型准确度较为理想，可靠性较强，但是将舆情倾向重新定义后，模型精度得到了较大提高，对于负面舆情的识别精度有了较大提升，本模型可以用于实际生活中车企舆情的判断。最后本文提出展望，在训练模型时数据选取时应尽量使得各类样本的数据占比均衡，避免造成过度识别问题。

关键词

车企舆情，朴素贝叶斯，舆情识别与预测

Recognition and Prediction of Positive and Negative Opinions of Car Companies

Zhuang Wu

College of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 1148322387@qq.com

Received: Dec. 25th, 2020; accepted: Jan. 19th, 2021; published: Jan. 26th, 2021

Abstract

With the continuous advancement of technology, people's lives are getting better and better, the

文章引用: 武壮. 车企舆情正负面识别与预测[J]. 计算机科学与应用, 2021, 11(1): 121-132.
DOI: 10.12677/csa.2021.111014

popularity of vehicles is gradually increasing, and people are paying more and more attention to the experience that vehicles bring to them. For auto companies, car safety is directly related to the lives of customers. People are more concerned about and sensitive to the positive and negative public opinions of auto companies, making it more difficult to deal with public opinions. If the negative public opinion cannot be dealt with in a timely manner, car companies will face significant public opinion pressure, and it will also consume a lot of resources and financial resources when dealing with it afterwards. Since most of the products are produced on a large scale and multi-party interests are entangled, the public opinion systems of auto companies often have higher public opinion requirements than other companies. Therefore, for auto companies, the identification and prediction of public opinion plays a very important role. The paper establishes a Naive Bayes Model to identify and predict the positive and negative public opinion of car companies. On the basis of effective data processing, this paper uses the given training set data to build the model, and uses the test set data to evaluate the rationality and scientificity of the model. Studies have shown that the accuracy and reliability of the public opinion recognition and prediction model for car companies adopted in this article is relatively satisfactory, but after redefining public opinion tendencies, the accuracy of the model has been greatly improved, and the accuracy of identifying negative public opinions has been greatly improved. This model can be used to judge the public opinion of car companies in real life. Finally, this article puts forward a prospect that when selecting data when training the model, we should try to balance the proportion of data of various samples to avoid over-identification problems.

Keywords

Public Opinion of Car Companies, Naive Bayes, Identification and Prediction of Public Opinion

1. 研究背景与意义

“舆情”在《现代汉语词典》(第七版)中的意思是指“公众的意见和态度”。通俗点理解，指的是群众对于社会中某些现象、事件甚至是问题所持有的态度和意见。车企舆情，顾名思义，就是大众对于汽车相关企业的行为或者是产品所表达的自己的意见和看法。

在如今网络信息越来越发达的时代，我国拥有世界上最多的网民和最大的网络访问量，以互联网为载体的网络舆情愈发活跃。网络不仅成为人们获取重要信息的重要渠道，也成为人们表达自己的观点和想法的平台。人们可以通过文字、图片、音频甚至是视频等多种形式进行沟通交流，网络舆情则成为了社会舆情主要的表现形式，具有直接、随意、突发、多元、偏差等特点，一旦处理不当，可能会对社会产生较大影响。因此，越来越多的专家学者甚至是企业投入到舆情的分析研究当中。

随着社会的发展，人们生活水平的提高，汽车已经成为许多家庭的必备产品。我国汽车市场也逐渐进入了供大于求的买方市场。汽车企业为了吸引顾客，适应市场态势，需要转换营销策略。通过对互联网上消费者对汽车品牌的看法与评论进行分析，了解消费者的购买行为和品牌认知等相关信息，可以很好得为企业在产品改进、竞争力优化等方面提供依据，对汽车企业应对新的市场需求有着较大帮助[1]。汽车安全直接关系消费者的生命安全，因此汽车行业的负面舆论较其他行业更容易引起关注度。然而汽车企业往往对于网络舆情实时监控力度不够，不能及时有效地控制舆论；对舆论的正负面识别，有助于汽车企业在危机公关处理问题上具有更强的针对性，提升企业形象。

2020 年伊始新冠肺炎疫情肆虐，汽车行业从生产到销售各个阶段都受到了不同程度的影响。由于交通物流受阻、员工无法到岗等情况，一些生产厂商被迫停产；经销商也面临着销量下滑、收入锐减的境况[2]。为了提升企业市场竞争力，各品牌厂商不仅重点关注产品的质量安全以及安全隐患，还致力于样汽车新科技等。汽车行业舆情热点涵盖范围广，企业在了解舆情主要倾向、知悉网民对本企业品牌态度之后，对于企业决策有一定的帮助。

2. 国内外研究现状

(一) 国外研究现状

在舆情分析方面，国外研究开始得要比我国早。Jeonghee Yi [2]等(2003)通过使用语法分析器和情感词典，从文档中提取与特定主体相关的正面或负面的情感，以正确识别情感表达与主题之间的语义关系，并且获得了很高的精度；Tetsuya Nasukawa 等[3] (2003)使用情感分析器从在线文档中提取出主题相关观点，并使用自然语言处理技术确定情感；YW Seo 等[4] (2004)采用 single-pass 算法对新闻进行聚类，发现该方法具有计算简单且运算效率高的特点；Hurtado J L [5] (2016)使用关联分析和整体预测从一组文本文档中自动发现主题，并预测其在将来的发展趋势。

(二) 国内研究现状

直到上世纪末期，我国才有学者开始对舆情进行分析研究[6]。许鑫等[7] (2008)分析了互联网舆情研究的现状，并通过支持向量机用于对网络舆情内容的主题聚类；王兰成等[8] (2013)将 HowNet 与主题领域语料的情感概念结合，并利用情感本体抽取特征词并判断其情感倾向度，结合句法规则及程度副词影响，采用机器学习的方法对主题网络舆情 web 文本进行倾向性分析；朱建平等[9] (2016)利用 2015 年第二季度中国房地产相关数据对房地产网络舆情进行了实证研究与分析，并且对发现的热点话题整体倾向性进行了评述。

文本倾向性分析是指挖掘出人们对于某件事物持有的态度或看法是正面还是负面。国内也有不少学者在这方面进行了研究。高洁等[10] (2004)讨论了朴素贝叶斯、K-邻近、支持向量机等常用的文本分类原理与方法；黄萱菁等[11] (2011)结合学术界近年文本情感分析的研究成果，对方法进行了概括归纳，并且对倾向性分类、倾向性分析应用等方面的研究现状进行介绍，最后还对情感倾向性分析技术进行了总结，展望了未来；许鑫等[12] (2011)尝试将基于统计和语义两种文本倾向性分析的方法结合起来，提出了基于模式抽取和匹配基础上的文本倾向性分类算法，并结合领域应用进行实证分析。

本研究通过朴素贝叶斯对互联网上车企相关舆情进行分析，对车企的成长与发展具有十分重要的意义与价值。

3. 研究内容框架

本研究主要分为三个部分，具体流程如图 1 所示。

第一部分：对文本数据进行预处理。本文给出的原始数据是未经过处理的数据，其中包含重复数据、缺失值、异常值等情况，故利用 Python 软件首先进行数据预处理，使得数据变成可直接使用的数据。在此基础上，利用 Python 的 jieba 库对文本数据进行分词处理，并提取相应特征。

第二部分：模型的建立。为了保证本文研究的科学性和合理性，首先利用训练集数据建立模型，之后用测试集数据对建立的模型进行验证。在模型建立过程中，本文拟采用朴素贝叶斯的方法，根据数据的文本特征将其分为正面、中性和负面三类。

第三部分：结果验证。将处理好的测试集数据带入建立好的模型中，通过实际的结果和模型的结果的比较，验证模型的合理性。

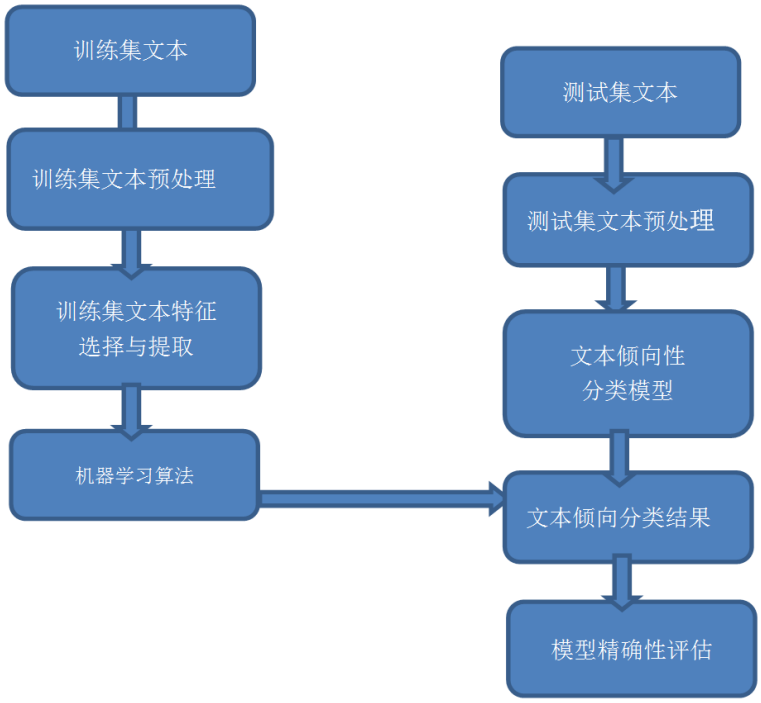

Figure 1. Flow chart of text orientation analysis
图 1. 文本倾向性分析流程图

4. 研究方法

(一) 朴素贝叶斯分类

1) 方法原理

目前国内外对于文本倾向性分析的方法主要有基于文本分类的文本倾向性分析、基于语义规则模式的文本倾向性分析和基于情感词的文本倾向性分析三大类。对文本进行分类的常见计算机分类器有 KNN (K-近邻法)、SVM (支持向量机)、NB (朴素贝叶斯)等。朴素贝叶斯是一种思想较为简单的方法，具有不错的鲁棒性，容易实现，运行速度快，因此被广泛使用[13]。

朴素贝叶斯(Naive Bayesian)是基于条件概率、贝叶斯定理和特征条件独立假设的分类方法，它通过特征计算分类的概率，选取概率大的情况进行分类[14]。它是一种十分简单的分类算法，朴素贝叶斯的思想基础是：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此分类项属于哪个类别。

朴素贝叶斯分类的正式定义如下：

设 $x=\{a_1,a_2,\cdots,a_m\}$ 为一个待分类项，而每一个 a 为 x 的一个特征属性。有类别集合 $c=\{y_1,y_2,\cdots,y_n\}$，计算 $p(y_1|x),p(y_2|x),\cdots,p(y_n|x)$。如果 $p(y_k|x)=\max\{p(y_1|x),p(y_2|x),\cdots,p(y_n|x)\}$，则 $x\in y_k$。

朴素贝叶斯最常见的分类应用是对文档进行分类，因此，最常见的特征条件是文档中出现词汇的情况，通常将词汇出现的特征条件用词向量 ω 表示，由多个数值组成，数值的个数和训练样本集中的词汇表个数相同。

朴素贝叶斯条件概率公式可表示为：

$$p(c_i|\omega)=\frac{p(\omega|c_i)p(c_i)}{p(\omega)} \tag{1}$$

如果 $p(c_1|\omega)>p(c_2|\omega)$，那么分类应当属于 c_1；如果 $p(c_2|\omega)<p(c_1|\omega)$，那么分类应当属于 c_2；

朴素贝叶斯方法有一个很重要的假设，就是基于特征条件独立的假设，也就是我们姑且认为词汇表中各个单词独立出现，不会相互影响，因此，$p(\omega|c_i)$可以将 ω 展开成独立事件概率相乘的形式，因此：

$$p(\omega|c_i)=p(\omega_0|c_i),p(\omega_1|c_i),p(\omega_2|c_i),\cdots,p(\omega_n|c_i) \tag{2}$$

2) 分类流程图

本文的贝叶斯分类流程分为三部分，如图 2 所示。

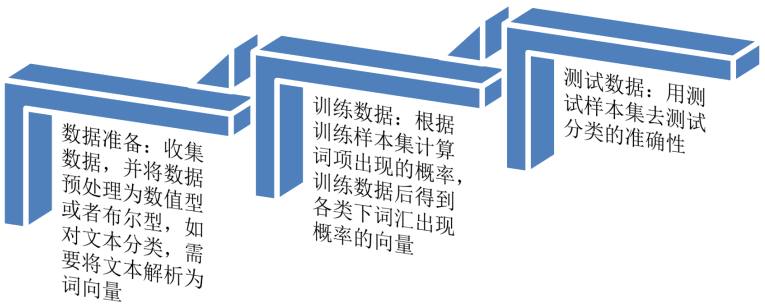

Figure 2. Naive Bayes flow chart
图 2. 朴素贝叶斯流程图

(二) 文本预处理

对文本数据进行预处理是为了选择合适的文本特征进行模型建立，以便于计算机能够准确识别并对数据进行处理。文本预处理的过程主要包括：文本分词、剔除停用词(包括没有实际意义的词语、颜文字和标点符号) [15]。

1) 文本分词

分词指是将中文长语句切割为一系列单独活动、结构最小的词。是文本分析中必不可少的一个过程，分词结果对于后续的文本特征提取以及文本倾向性分析都会产生重要影响。

jieba 因其安装简单，且有精确、全和搜索引擎三种模式，支持简体、繁体中文，受到广泛使用。并且 jieba 库还具有词性标注功能，可以标注句子分词后每个词的词性，词性标注集采用北大计算所词性标注集，属于采用基于统计模型的标注方法[16]。

2) 剔除停用词

在分词之后，还需要将一些分辨能力差或根本没有分辨能力的词语(如“的，了”)，即不能传达情感的词语过滤掉。这些词通常指介词、连词以及一些英文单词、数字、标点符号等[15]。

本研究综合采用了哈工大停用词表、百度停用词表、四川大学机器职能实验室停用表等四个停用表，尽可能提高保留下的文本数据有效性。

(三) 数据质量评估

数据质量评估是提高数据质量的基础和必要前提，它能对应用数据的整体或部分数据的质量给出一个合理的评估，从而帮助用户了解数据质量水平，并采取相应措施以提高数据质量[17]。

数据质量的衡量指标主要包括数据的准确性、完整性、一致性、有效性、覆盖率等。即检验数据是否与其描述主题保持一致，数据是否存在缺失记录或字段，描述统一实体相同属性的值在不同数据集中是否一致，数据是否满足使用条件，是否含有不合法字段或不规则数据，是否存在重复记录；数据来源广度如何、覆盖的人群、地点等等是否符合数据要求[18]。

5. 实证分析

(一) 数据的预处理和数据质量评估

1) 数据来源

本文中的数据来源于全国第四届应用统计案例大赛案例 C：车企舆情正负面识别与预测。数据结构包括文章题目，文章内容，文章来源网站名称，文章网址和已经人工标注的舆论倾向，数据集已经被分为了训练集和预测集。

2) 数据预处理

通过所提供数据的数据结构，可以推测该数据是在疫情期间从车企相关网站上通过关键词进行爬虫获取的。从数据结构上推测，原始数据的第一列为获取文章的标题，本文将其命名为“title”，第二列为文章的主体，本文将其命名为“passage”，第三列为文章来源的网站名称，本文将其命名为“website”，第四列为文章的互联网协议地址，本文将其命名为“http”，最后一列是则是文章的舆论倾向变量，其中“1”代表该文章对车企具有正向舆论导向，“−1”表示该文章包含对于车企的负面评论，“0”代表该文章并不包含有关车企正面或鼓面的舆论，对于车企没有明确的情感倾向。

在网络爬虫中，由于网络信息的复杂性，可能造成所爬取的文本信息存在一定的问题，因此需要对爬取的数据进行一定的预处理。数据预处理共分以下几个步骤。

① 案例中所给数据为 csv 格式，由于 csv 格式的文件是由逗号分隔，容易与文章的逗号混淆，因此需要先转存为 excel 文件，共得到 99,842 条数据。

② 对于某些条目含有空值的数据，也进行删除处理，共删除含有空值的数据 136 条，剩余数据 99,706 条。

③ 数据中有大量重复冗余数据，因此通过比对数据的文章标题，对于文章标题相同的多条数据，本文仅保存一条，删除其他数据。经过本步处理后还剩余 69,825 条数据。

④ 文章内容中包含了大量的网址和数字，这对正负面舆论的识别并没有显著的影响，因此运用正则表达式对文章中的网址和数字进行识别和剔除。

⑤ 最后，本文认为低于 10 个字符的文章对于车企正负面舆情的识别参考意义较小，因此将文章长度小于 10 个字符的数据进行删除，得到最终的数据共 68,060 条。

3) 数据质量评估

① 数据完整性。训练集数据的数据完整性较好，在得到的 99,842 条数据中，仅有 136 条数据包含空值，有值率达到 99.36%。但是，数据集并没有给出每一列数据所代表的具体含义，这对于数据的正确理解分析会造成一定的偏差。

② 数据有效性。借助 Python 编程对训练集数据进行全局去重发现，数据集中含有 29,881 条重复冗余数据，冗余率高达 30.00%。同时，文章字段数少于 10 个字符的可以认定为无效数据，这样的数据共有 1765 条，综上所述，有效数据仅占 68.26%，这其实也是由于文本型数据的获取方式和复杂性所导致的。

③ 数据覆盖率。经过处理后的训练集数据来源广泛，不仅千里马、广元汽车之家、伯乐二手车网等汽车行业媒体，而且包含新浪网、东方财富网、今日关注这些财经类、社会关注类网站和一些自媒体。

(二) 训练集数据的描述统计分析

1) 车企舆论倾向分布描述

使用 python 对处理后的车企舆情训练集数据舆论倾向分布进行统计，结果发现正面舆论新闻有 9720 条，中性舆论 49,186 条，负面舆论 9154 条，分布结果如图 3 所示：

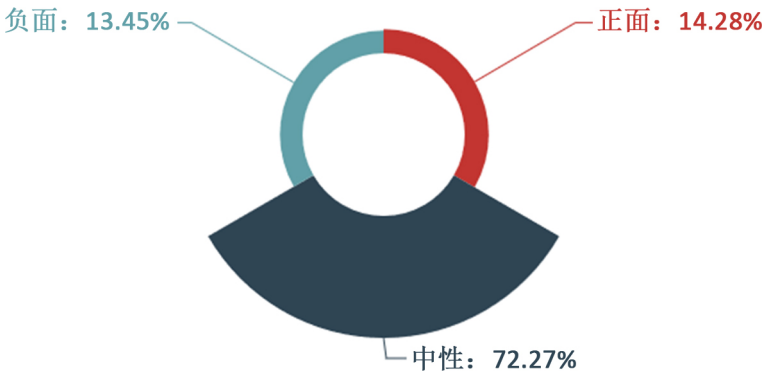

Figure 3. The distribution of public opinion tendency of car companies in the training set data
图 3. 训练集数据车企舆论倾向分布

在该分布图中，环形的宽度表示各舆情倾向的占比。可以看出，大部分网络文章对于车企行业没有明显的舆情倾向，持负面和正面舆情倾向的文章占比分别为 13.45%和 14.28%。车企可根据正负面舆情的相关方面，针对性地改进完善产品及服务。

在所给的训练集数据中，由于绝大部分文章没有明显的舆论倾向，因此在构建词频矩阵的时候，中性舆情文章的词汇特征会被较多提取，在进行模型学习时可能会造成中性舆情文章识别过度的问题，从而导致具有明显情感倾向的文章不容易被识别出来。

2) 分网站车企行业舆情倾向分布描述

我们还选择了数据来源最多的四家网站，以分析舆论倾向是否与来源网站有关。处理过后的训练集数据中有 1370 条来自“千里马”网站，通过图 4 可以看到，该网站舆论消息与总体舆论倾向基本保持一致，仍是以中性舆论居多；不同的是，负面舆论 115 条，占总数的 8.39%；正面舆论只有 32 条，占比为 2.34%。

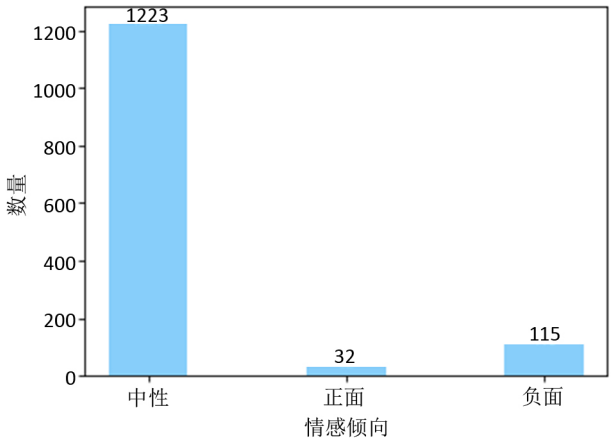

Figure 4. The distribution of public opinion tendencies of auto companies on the “Qianlima” website
图 4. “千里马”网站车企舆论倾向分布

舆论来源第二多的网站是“广元新媒体”，共 726 条舆论数据。由图 5 可知，该网站的车企舆论分布倾向与总体分布倾向有较大的差异。其中，正面舆论 484 条，占比 66.67%超过总数的一半；中性舆论只有 37 条，占总数 5.10%，可以看出该网站的文章对于车企主要持正面态度。

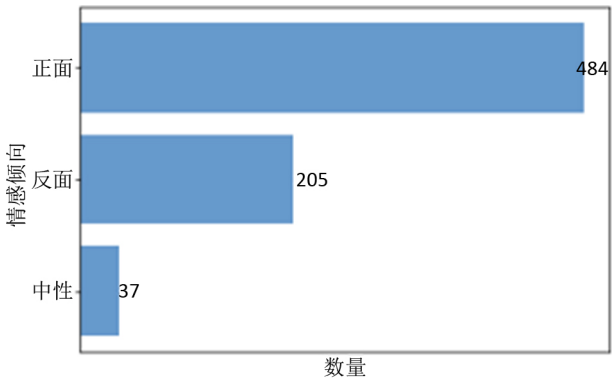

Figure 5. Distribution of public opinion tendencies of auto companies on the "Guangyuan" website
图 5. "广元新媒体"网站车企舆论倾向分布

选取四家车企舆论数量最多的网站进行对比，通过图 6，我们发现：除"广元新媒体"以外，其他三家网站均是中性车企舆论最多，且占比均超过网站车企舆论总数的 65%；"千里马"网站的负面舆论数量则相比正面舆论要多，其余三家则都是正面舆论数量多于负面，可以看出不同网络来源的文章之间的舆论倾向分布具有较大差异，为进一步验证本文的观点，需要做简单的方差分析进行检验。

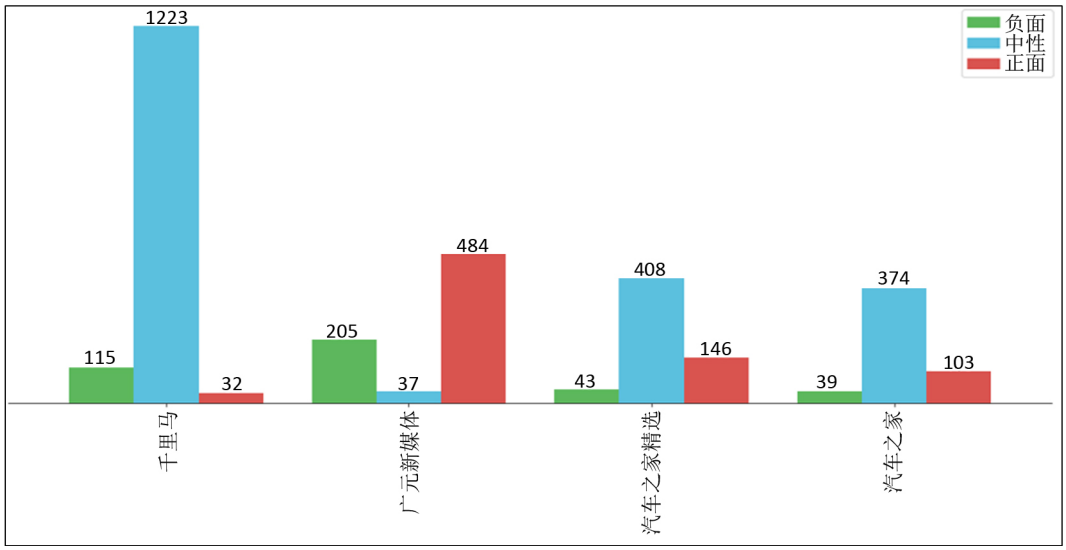

Figure 6. The distribution of public opinion tendencies of auto companies on the "Qianlima" website
图 6. "千里马"网站车企舆论倾向分布

用 python 对四个网站的舆论分布进行单因素方差分析，得到结果如表 1 所示。由表 1 可知方差分析的 F 统计量为 67.9009，相伴概率 P 值远远小于显著性水平 0.05，因此拒绝不同网站的舆论倾向分布没有差异的原假设，认为车企舆论分布倾向与网站来源有关，因此在进行分词时应把网站名称加入到文章中。

Table 1. Variance analysis results of the distribution of public opinion tendencies of various websites
表 1. 各网站车企舆论倾向分布方差分析结果

	df	sum_sq	mean_sq	F	PR (>F)
group	1.0	23.3449	23.34489	67.9009	2.48×10^{-16}
Residual	3207.0	1102.5928	0.3438		

3) 分舆论倾向的主要词汇特征描述

Figure 7. Word cloud diagram of positive public opinion tendencies
图 7. 正面舆论倾向词云图

从图 7 可以看出，在给出的 68,060 条评论中，轿车和公司是所有评论中出现频次最高的两个词，除此之外，库里南是正面舆论倾向中频次出现最多的词，由于其本身安全性高、抗震性强、品质高等特点，受到人们一致好评。从词云图可以看出，正向舆论倾向中，出现频次较高的词更多是与汽车品牌有关，例如：“库里南”、“马自达”、“起亚”等。

Figure 8. Word cloud diagram of negative public opinion tendencies
图 8. 负面舆论倾向词云图

由图 8 可知，在给出的评论中，“汽车”、“车型”、“帕萨特”、“成绩”、“碰撞”、“本田”等成为了负面舆论倾向中的关键词云，在相关信息中被提及频次较高。“碰撞”在负面舆情词云图中出

现频次很高，说明其对于车企或者消费者而言，是社会关注度和敏感度较高的词，且其舆论导向性很高；对比正向舆论词云图，车型提及频次高很多，由此可以看出，消费者在汽车外形方面，更注重车型，而大多数车辆的车型并不能很好地满足消费者的购车需求。除此之外，看整个词云，可以看出，负面舆论倾向中，更多的评论是与汽车本身性能和性价比有关，“气囊”、“低速”、“低配”、“性价比”、“费用”、“伤害”等出现频繁，很多车对于消费者而言性价比等并不高，不受消费者的青睐。

Figure 9. A word cloud diagram of neutral public opinion tendencies
图 9. 中性舆论倾向词云图

由图 9 可知，在所有评论中，“市场”、“全球”、“车辆”、“女司机”、“广州”等词成为了中性舆论倾向中出现频次较高的词，结合整个中性舆论词云图可以看出，与中性舆论关联性较高的词基本属于宏观方面，大多是从社会整体的角度发表评论，没有任何舆论导向。

(三) 朴素贝叶斯模型的训练与测试

通过对已经预处理好的数据进行分词、去除停用词，我们得到了文本数据的词频矩阵。通过训练集数据的词频矩阵构建朴素贝叶斯分类器，然后将测试集的数据放入分类器中进行预测，从而对分类器的稳健性和预测效果进行评估。测试集车企舆情预测精度如表 2 所示。

Table 2. Test set accuracy
表 2. 测试集精度

Values	Precision	Recall	F1-Score	Support
−1	0.45	0.68	0.54	8992
0	0.86	0.68	0.76	49850
1	0.37	0.58	0.45	9818
Accuracy			0.67	68660
Macro Avg	0.56	0.65	0.59	68660
Weighted Avg	0.73	0.67	0.69	68660

从该结果中可以看到，车企负面舆情的识别精度为 0.45，召回率为 0.68，F1 得分为 0.54；车企中性

舆情的识别精度为0.86，召回率为0.68，F1得分为0.76；车企正面舆情的识别精度最低，仅为0.37，召回率为0.58，F1得分为0.45。从全局来看，车企舆情识别的总体精度为0.67，按照数据量加权计算后的识别精度为0.73，还是比较理想的。

造成识别精度差的原因主要有以下几点：1) 数据分布不均衡，正面和负面舆情的关键词汇获取较少；2) 正面和中性舆情倾向的文章词汇相似度较高，不易区分；3) 文章较长，词汇矩阵过大导致的预测精度降低。

(四) 数据集的重新分类与验证

考虑到正面和中性舆情倾向的文章不容易区分，本文拟重新定义舆情倾向编码，并进行分类预测。

1) 负面舆情导向与其他

本文将车企的正向舆情倾向与中性舆情倾向归为一类，记为“0”，将负面舆情倾向记为“−1”，重新利用训练集数据进行朴素贝叶斯分类器的构造，然后将训练集数据放入分类器中得到的精度报告如表3所示：

Table 3. Test set accuracy (0 VS −1)
表3. 测试集精度(0 VS −1)

Values	Precision	Recall	F1-Score	Support
−1	0.42	0.70	0.53	8992
0	0.95	0.86	0.90	59668
Accuracy			0.84	68660
Macro avg	0.69	0.78	0.71	68660
Weighted avg	0.88	0.84	0.85	68660

从该结果中可以看到，车企负面舆情的识别精度为0.42，召回率为0.70，F1得分为0.53；车企非负面舆情的识别精度为0.95，召回率为0.86，F1得分为0.90。从全局来看，车企舆情识别的总体精度为0.84，按照数据量加权计算后的识别精度为0.88，与未重新定义之前，精度有了很大的提升。

2) 正面舆情导向与其他

本文将车企的负向舆情倾向与中性舆情倾向归为一类，记为“0”，将正面舆情倾向记为“1”，重新利用训练集数据进行朴素贝叶斯分类器的构造，然后将训练集数据放入分类器中得到的精度报告如表4所示：

Table 4. Test set accuracy (0 VS 1)
表4. 测试集精度(0 VS 1)

Values	Precision	Recall	F1-Score	Support
0	0.92	0.83	0.87	58842
1	0.36	0.59	0.45	9818
Accuracy			0.79	68660
Macro avg	0.64	0.71	0.66	68660
Weighted avg	0.84	0.79	0.81	68660

从该结果中可以看到，车企正面舆情的识别精度为0.36，召回率为0.59，F1得分为0.45；车企非正面舆情的识别精度为0.92，召回率为0.83，F1得分为0.87。从全局来看，车企舆情识别的总体精度为0.79，按照数据量加权计算后的识别精度为0.84，与未重新定义之前，精度有了很大的提升。但是，与第一种定义方式相比，精度有所下降，也从侧面印证了负面舆情比正面舆情更具有识别性。

6. 结论与展望

(一) 模型结论

1) 本文利用朴素贝叶斯分类模型对文本型数据进行舆情倾向识别，加权识别精度达到 0.73。

2) 本文将舆情倾向重新编码为负面舆情及其他和正面舆情及其他，重新编码后的识别精度相比之前有了较大提升，加权识别精度分别达到了 0.88 和 0.84，并且本文发现负面舆情相比正面舆情更容易识别。

3) 车企应更多地关注负面舆情相关文章，进行词云分析，以找到消费者的迫切需求和需要提升的具体方面。

(二) 模型展望

1) 文本型数据具有一定的复杂性，在本文中，直接借用了已有研究的停用词列表，之后应该根据实际数据的词频统计等相关信息，构建自己的停用词列表。

2) 本文利用分词结果构建词频矩阵进而对文章的舆情倾向进行分类，然而，有些词汇可能大量出现在各种倾向的文章中，这些词汇本身没有明显的舆论倾向，但是组合成一定的词组和短句后便具有了明显的舆情倾向，因此利用词组和短句构建预测矩阵也将是以后需要研究的方向。

3) 文中所给的数据中，没有舆情倾向的中性数据占比较大，造成中性舆情的过度识别，在进行训练集数据选取时应尽量使得各类样本的数据占比均衡，避免造成过度识别问题。

4) 本文运用朴素贝叶斯分类方法对文本数据进行分类，该方法要求各样本之间互相独立，对于大量数据的预测效果可能不佳，本文应该进一步采取其他分类方法加以比较。

参考文献

[1] 贺畅, 赵威, 陈陌. 基于网络舆情分析的汽车市场及消费研究[J]. 汽车工业研究·月刊, 2016(4): 4-9.

[2] Nasukawa, T. and Yi, J. (2003) Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In: *Proceedings of the 2nd International Conference on Knowledge Capture*, ACM, New York, 70-77. https://doi.org/10.1145/945645.945658

[3] Yi, J., Nasukawa, T., Bunescu, R., *et al*. (2003) Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques. *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, 19-22 December 2003, 427-434.

[4] Seo, Y.W. and Sycara, K. (2004) Text Clustering for Topic Detection.

[5] Hurtado, J.L., Agarwal, A. and Zhu, X. (2016) Topic Discovery and Future Trend Forecasting for Texts. *Journal of Big Data*, **3**, 1-21. https://doi.org/10.1186/s40537-016-0039-2

[6] 陶慧. 网络舆情热点话题发现研究[D]: [硕士学位论文]. 厦门: 厦门大学, 2017.

[7] 许鑫, 章成志. 互联网舆情分析及应用[J]. 情报科学, 2008, 26(8): 1194-1204.

[8] 王兰成, 徐震. 基于情感本体的主题网络舆情倾向性分析[J]. 信息与控制, 2013(1): 46-52.

[9] 朱建平, 等. 中国房地产网络舆情分析[J]. 数理统计与管理, 2016, 35(4): 722-741.

[10] 高洁, 吉根林. 文本分类技术研究[J]. 计算机应用研究, 2004, 21(7): 28-30.

[11] 黄萱菁, 等. 文本情感倾向分析[J]. 中文信息学报, 2011, 25(6): 118-126.

[12] 许鑫, 等. 一种文本倾向性分析方法及其应用[J]. 现代图书情报技术, 2011(10): 54-62.

[13] 朴素贝叶斯算法原理及实现[EB/OL]. https://www.cnblogs.com/sxron/p/5452821.html

[14] 机器学习之朴素贝叶斯(NB)分类算法与 Python 实现[EB/OL]. https://blog.csdn.net/moxigandashu/article/details/71480251, 2017-05-09.

[15] 范建美. 中文短文本情感倾向性分析研究[D]: [硕士学位论文]. 石家庄: 河北科技大学, 2015.

[16] jieba 分词原理[EB/OL]. https://blog.csdn.net/sinat_26811377/article/details/100703439

[17] 杨青云, 赵培英, 杨冬青, 唐世渭, 童云海. 数据质量评估方法研究[J]. 计算机工程与应用, 2004(9): 3-4+15.

[18] 如何评估数据质量[EB/OL]. https://www.cnblogs.com/hejunhong/p/12000216.html